



TCI-UNet: transformer-CNN interactive module for medical image segmentation

XUAN BIAN, GUANGLEI WANG,* YANLIN WU, YAN LI, AND HONGRUI WANG

College of Electronic and Information Engineering, Hebei University, Hebei 071002, China

**513197133@qq.com*

Abstract: Medical image segmentation is a crucial step in developing medical systems, especially for assisting doctors in diagnosing and treating diseases. Currently, UNet has become the preferred network for most medical image segmentation tasks and has achieved tremendous success. However, due to the limitations of convolutional operation mechanisms, its ability to model long-range dependencies between features is limited. With the success of transformers in the computer vision (CV) field, many excellent models that combine transformers with UNet have emerged, but most of them have fixed receptive fields and a single feature extraction method. To address this issue, we propose a transformer-CNN interactive (TCI) feature extraction module and use it to construct TCI-UNet. Specifically, we improve the self-attention mechanism in transformers to enhance the guiding ability of attention maps for computational resource allocation. It can strengthen the network's ability to capture global contextual information from feature maps. Additionally, we introduce local multi-scale information to supplement feature information, allowing the network to focus on important local information while modeling global contextual information. This improves the network's capability to extract feature map information and facilitates effective interaction between global and local information within the transformer, enhancing the representational power of transformers. We conducted a large number of experiments on the LiTS-2017 and ISIC-2018 datasets to verify the effectiveness of our proposed method, with DCIE values of 93.81% and 88.22%, respectively. Through ablation experiments, we proved the effectiveness of the TCI module, and in comparison with other state-of-the-art (SOTA) networks, we demonstrated the superiority of TCI-UNet in accuracy and generalization.

© 2023 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

1. Introduction

The convolutional neural network (CNN), especially the fully convolutional network (FCN) [1], has dominated medical image segmentation, and UNet [2] has achieved remarkable results. It consists of a symmetrical encoder-decoder network and complements feature map details through skip connections. Since then, improved methods based on this encoding-decoding structure have continuously been proposed, and have achieved great success in medical applications such as heart segmentation in magnetic resonance, organ segmentation in CT, and polyp segmentation in colonoscopy videos, such as UNet++ [3] and UNet3+ [4], which enhanced the feature information passed between the encoder and decoder by improving the skip connection structure and relieved the semantic gap problem. H-DenseUNet [5] introduced dense connection structure into the network inspired by DenseNet [6] and achieved liver segmentation. Attention UNet [7] introduced the idea of attention mechanism on top of the encoder-decoder network, improving the network's ability to extract important features. CENet [8] added multi-scale mechanism to the bottleneck part of the network, enhancing the network's ability to handle deep feature map semantic information.

Although CNN-based methods have outstanding performance, they lack the ability to model long-distance dependencies between features due to the inherent limitations of the convolution operation's receptive field. Transformer [9], initially used to model sequence-to-sequence prediction in NLP tasks, has recently gained great interest in the field of computer vision. After ViT [10] introduced Transformer to the field of computer vision in 2020, it solved the limited receptive field problem of CNN. Transformer maps feature maps to three output feature vectors, Query (Q), Key (K), and Value (V), and allocates attention resources to V through the matrix multiplication of Q and K to obtain an attention map, using the multi-head self-attention mechanism. Transformer excels in modeling global context, but it has limitations in capturing fine-grained features.

Both CNN and Transformer-based models have strengths and weaknesses in feature extraction. CNNs excel at capturing local information, but lack the ability to model long-range dependencies between features. On the other hand, Transformers excel at handling global context information, but lack attention to the connections and influence between local features, especially when small target regions need to be segmented, where local dependencies are more important than global context. Researchers have combined the advantages of both by combining CNNs and Transformers, such as TransUNet [11], which first extracts low-level features with CNNs and then simulates global interactions with Transformers. FAT-Unet [12] is a network used for skin lesion diagnosis, which adopts a dual-branch encoder with convolutional and Transformer encoders. TransFuse [13] re-designs the encoder and uses both convolutional and Transformer encoder branches to process the output feature map. However, we find that previous work on combining CNNs and Transformers has the following shortcomings: (1) The main focus in terms of structure is on combining a simple Transformer with a CNN encoder-decoder, lacking the ability to interact between local and global information inside the Transformer. (2) The amount of information contained in the Q and K vectors of the Transformer is still large and the computational cost is still high, and the attention maps generated by these vectors still have room for further improvement in guiding resource allocation. (3) Most existing Transformer networks have a single receptive field size and are unable to adapt to the characteristics of medical image multi-scale changes, lacking the ability to model multi-scale information.

To address these problems, this paper presents a Transformer-CNN interaction (TCI) feature extraction module, as shown in Fig. 1. (a) is the original Transformer structure diagram, and (b) is the improved TCI module structure diagram. The specific improvement is to replace the multi-head self-attention (MSA) in the original Transformer structure with the TCI module proposed in this paper. This module implements the interaction between local and global information inside the Transformer, reduces the computational cost by reducing the dimensionality of the feature information of Q and K vectors, and further strengthens the network's ability to guide resource allocation through channel attention mechanisms. A pyramid structure is designed to provide rich multi-scale information for the self-attention mechanism and to enrich the receptive field of each layer of the feature extraction structure, better adapting to the medical image segmentation task.

In conclusion, our contribution can be summarized as:

1. A Transformer-CNN interactive (TCI) module is proposed, combining the advantage of Transformer capturing long-range dependencies with the inductive bias advantage of Convolutional Neural Networks. The improved self-attention mechanism and carefully designed convolutional structure interact local information and global contextual information within the Transformer, and a TCI-UNet is constructed for medical image segmentation.
2. A Window Self-Attention Enhance (WSE) block is proposed. The block adopts a windowed approach to restrict the global self-attention mechanism to a local region. By reducing the dimensions of Q and K along the channel dimension, the computation complexity of the model is reduced. The channel attention mechanism is then used to filter and enhance the

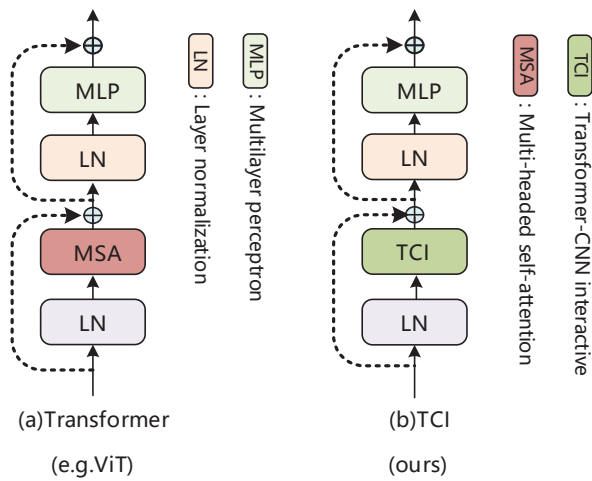


Fig. 1. Structure of Transformer and TCI Module. (a) shows the structure diagram of Transformer, while Fig. 1 (b) shows the structure diagram of TCI module proposed in this paper. TCI has been re-designed and improved at the position of MSA, realizing the interaction between local information and global information inside the Transformer.

feature information of Q and K, thereby enhancing the attention map's ability to guide the allocation of computational resources and improving the model's accuracy.

3. A Multiscale Group Convolution (MGC) block is proposed, which employs a pyramid structure to capture multi-scale local information. Rich spatial local information is mined in three paths with different receptive fields. Multiple convolution kernels map local information from small to large, extracting information at different scales, achieving multi-scale representation of the local information in the feature map, and utilizing group convolution form to compress computation and reduce the computational burden introduced into the network.

2. Related work

2.1. Transformers in the field of computer vision

Transformer was initially proposed in the NLP field and quickly became the dominant architecture in this field. In 2020, ViT introduced the Transformer into the computer vision field, dividing the input image into 16×16 patches and mapping them into two-dimensional token vectors through an embedding layer. The token vectors, after being encoded with position information, were then processed in the Transformer architecture.

With the success of ViT in computer vision, researchers have proposed many improved Transformer models. For example, Swin Transformer [14] built a hierarchical structure similar to convolutional neural networks and optimized the self-attention mechanism through the combination of Windows Multi-Head Self-Attention and Shifted Windows Multi-Head Self-Attention. Similarly, HaloNet [15] also achieved results surpassing convolutional neural networks on the ImageNet dataset through a hierarchical architecture and window self-attention mechanism. Additionally, CSWin Transformer [16] proposed a Cross-Shaped Window self-attention (CSwin Self-Attention) mechanism, which separately calculated self-attention in vertical and horizontal windows. It also introduced a form of local enhanced positional encoding to improve the positional encoding process in Transformer, thereby enhancing the model's competitiveness. In contrast, Neighborhood Attention Transformer [17] proposed a Neighborhood Attention mechanism that

adaptively locates the receptive field around each token without requiring additional operations. VSA [18] proposed a Varied-Size Window Attention to learn an adaptive window configuration from data. It used a window regression module to predict the target window's size and position, capturing rich context information from different windows and promoting information exchange between windows. BOAT [19] started from the image content, clustering tokens with similar features, and then calculated them using self-attention. While using local self-attention, it basically preserved the original global self-attention mechanism's ability to model long-distance dependencies between features.

Besides limiting the scope of attention computation using the window idea, some researchers improved Transformer by adopting the multi-path idea. For example, Inception Transformer [20] split the input feature map along the channel dimension and inputted the split feature maps into high-frequency and low-frequency mixers for feature extraction, combining the advantages of convolution and max pooling with Transformer. MPViT [21] built a multi-path visual transformer and simultaneously achieved fine and rough feature representations at the same feature level, resulting in excellent performance. Multiscale Vision Transformers [22] combined the idea of multi-scale features with multi-head pooling attention, expanded the channel dimension in a hierarchical manner and reduced the spatial resolution, effectively combining Transformer with multi-scale features. Despite the progress made by Transformer in the field of computer, there were still problems such as a lack of inductive bias, high computational complexity, a large amount of data required for training, and a certain degree of dependence on pre-training weights, which limited the model's flexibility.

2.2. Medical image segmentation network

Since UNet became the standard network for medical image segmentation tasks, many improved networks based on UNet have emerged, such as UNet++ and UNet3+ which were improved from the perspective of skip connections; U2Net [23], R2UNet [24], CENet, and MultiResUNet [25] which were improved from the network structure perspective; AnatomyNet [26], Attention U-Net, and RAUNet [27] which added attention mechanisms, etc. With the successful application of Transformer in the field of computer vision, researchers have also started to focus on the combination of UNet and Transformer structures. For example, TransUNet, which first introduced Transformer into the UNet structure, embeds the Transformer module into the encoder to extract global information of features. Influenced by the Swin Transformer, SwinUNet [28] abandoned the traditional convolutional architecture and built the encoding and decoding structure based on the Swin Transformer, which has been proved to have good segmentation accuracy and generalization ability through experiments. Similarly, DS-TransUNet [29], which is also built on the Swin Transformer, uses two independent branches with different patch sizes for multi-scale feature representation, enhancing the ability of the network encoder to represent both coarse-grained and fine-grained features. TransAttUNet [30] adds a novel self-attention and global spatial attention in the bottleneck part of the network, effectively learning non-local interactions among encoder features and aggregating features at different semantic scales with additional multi-scale skip connections. TransFuse re-designs the encoder by parallel processing the output feature maps with convolutional encoding branches and Transformer encoding branches, improving the modeling efficiency of the global context without losing low-level spatial information. Other networks with the dual-branch encoder of convolution and Transformer include FAT-Net for skin lesion segmentation and TUNet [31] for pancreas segmentation. Furthermore, UCTransNet [32] unifies the handling of different-scale skip connections using the Transformer structure, relieving the semantic gap problem caused by the original skip connection structure and achieving accurate medical image automatic segmentation. ScaleFormer [35] takes a scale-based approach to reexamine Transformer-based backbone networks, in order to adapt to the special requirements of multi-scale and high-resolution images in medical image segmentation tasks. TransDeepLab

[36] is based on the encoder-decoder Deeplabv3 + model and utilizes Swin-Transformer as its key component to explore the potential of Transformers for medical image segmentation.

Although the above-mentioned network structures have achieved certain achievements in the field of computer vision, they have not integrated the advantages of windowed self-attention mechanism, multi-scale thinking, and global and local information organically. Therefore, the performance limit of the model has not been further developed. Based on this idea, we have explored the unification of the three and designed the TCI module to improve the performance of the network model.

3. Methods

3.1. Overall architecture

The TCI-UNet introduced in this paper is built based on TCI module as the basic unit and follows a symmetrical U-shaped encoder-decoder structure, as shown in Fig. 2 (a). The network consists of four parts: an encoder, a bottleneck layer, a decoder, and skip connections. In the encoder, the image undergoes four-fold down sampling in the spatial dimension through Patch Partition and Linear Embedding and is compressed from $R^{C \times H \times W}$ to $R^{C \times L}$ form required by the Transformer. Patch Merging downsamples the input image three times, and the TCI module extracts local and global features of the different resolution feature maps. As the network deepens and the spatial resolution of the feature map decreases, the number of channels increases. The three skip connections transfer the corresponding resolution encoding information to the decoder to enhance the feature representation at each layer. In the decoder, Patch Expanding upsamples the input image processed by the bottleneck layer three times, and the TCI module processes information from both the skip connections and the upsampling path, gradually restoring the spatial resolution and mining local and global feature information. Finally, the feature map is upsampled four times to restore it to the original picture resolution. The entire network realizes end-to-end segmentation of medical images and balances the number of parameters and the accuracy of the network.

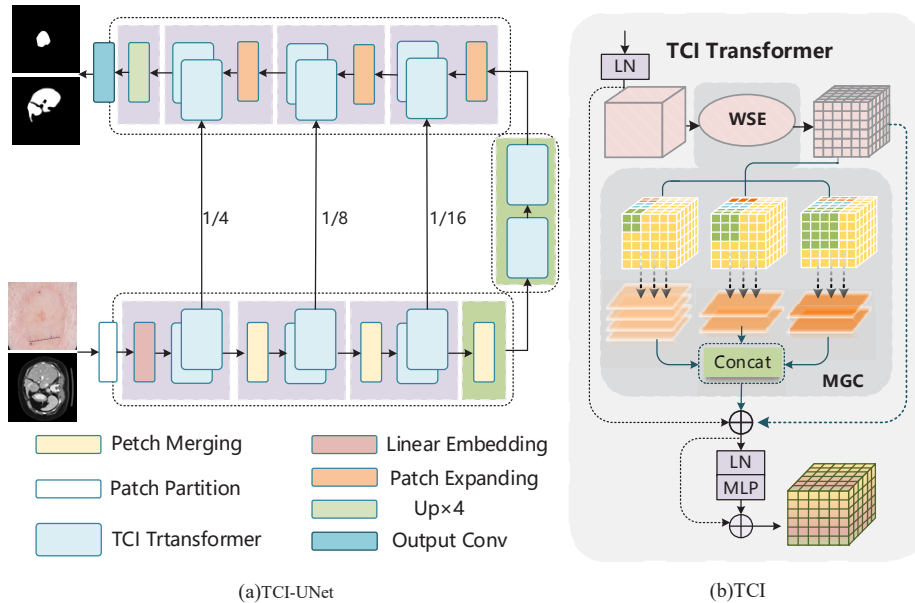


Fig. 2. Overall Framework Diagram. (a) TCI-UNet Structure; (b) TCI module Schematic Diagram.

The structure diagram of the TCI feature extraction module is shown in Fig. 2 (b). Its core part consists of two blocks, WSE and MGC. The WSE block focuses on mining global information, while the MGC block focuses on mining local information at multiple scales. The global information obtained after processing by the WSE block is input into the MGC block for further detailed information capture, and then the output result of the MGC block is combined with the complete global context information preserved by the skip connection, modeling the interaction between the global context information and the multi-scale local information inside the Transformer, which enhances the representation ability of the Transformer and improves the way of feature map information extraction of the network.

3.2. TCI module

The core of the traditional Transformer architecture consists of Multi-headed Self-Attention (MSA) and Multi Layer Perceptron (MLP). Layer normalization (LN) and residual connections are used to enhance the stability of feature distribution and gradient flow during network training. The output of the L-th layer Transformer in the network can be defined as:

$$S_L = \text{MSA}(\text{LN}(S_{L-1})) + S_{L-1}, \quad (1)$$

$$S_{L+1} = \text{MLP}(\text{LN}(S_L)) + S_L, \quad (2)$$

where S_{L-1} is the output feature map of the previous layer, S_L is the feature map calculated after self-attention mechanism, and S_{L+1} is the output feature map of this layer Transformer. As one of the core structures of Transformer, self-attention takes on the responsibility of exploring the global context information in the spatial dimension. The calculation formula is shown as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (3)$$

where Q , K , and V represent the Query vector, Key vector, and Value vector respectively, these three vectors are all mapped from the same input vector. The dot product of Q and K^T represents the exploration of attention in the spatial dimension, and the calculation result is the attention weight map. d_k is the number of channel dimensions of Q , and its main purpose is to control the range of the matrix values after Q and K are multiplied, to ensure the stability of the gradient during the *Softmax* operation. Finally, it is multiplied with V that represents the input feature map, realizing the application of the self-attention mechanism in the spatial dimension of the feature map.

Although this approach has made performance breakthroughs, the Transformer architecture based on self-attention mechanism is greatly affected by the input image resolution, especially for high-resolution images and dense tasks, which results in an unbearable computational cost. In addition, during the self-attention calculation process, the Query and Key vectors contain a lot of information to be processed, and some of the channels have little effective spatial information but a lot of invalid spatial information, which leads to the high computational cost of Self-Attention and weak attention guidance, greatly limiting the application of the Transformer structure in the field of medical image segmentation.

Inspired by CSwin Self-Attention, we proposes a new block called WSE to address the problems mentioned in the previous question. The block reduces the computational cost while improving the precision of attention mechanism. It does this by compressing the information to be processed by reducing the dimension of Q and K along the channel dimension and using a lightweight channel attention mechanism to filter the information. Additionally, the paper introduces MGC block that uses multi-scale group convolution to extract local information. The WSE block and MGC block together compose the TCI module. The TCI module combines the advantages of window self-attention mechanism and multi-scale convolution structure to model

the interaction between global context information and multi-scale local information. The overall structure of the TCI module is shown in Fig. 3, where the WSE block is responsible for exploring global context information and the MGC block is responsible for exploring multi-scale local information, and the output is combined with the preserved complete global context information through skip connections.

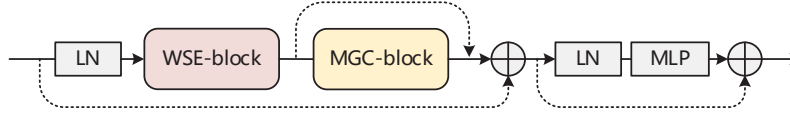


Fig. 3. Overall structure of TCI module.

3.2.1. Window self-attention enhance block

The figure for the WSE block is shown in Fig. 4. In it, Dimension decay convolution is a dimension decay convolution operation. For the input feature map $X \in R^{C \times L}$, three feature vectors ($Q \in R^{C \times L}$, $K \in R^{C \times L}$, $V \in R^{C \times L}$) are obtained through the linear mapping operation. Then, reshape Q and K and use a 3×3 convolution to reduce the channel dimension to 1/4 of its original size, resulting in $Q \in R^{\frac{C}{4} \times H \times L}$ and $K \in R^{\frac{C}{4} \times H \times L}$. Afterwards, ECA [33] is used in the channel dimension to enhance and select information. ECA is a lightweight structure and its formula is given as follows:

$$N = \delta(\text{Conv1d}(\text{GAP}(x))), \quad (4)$$

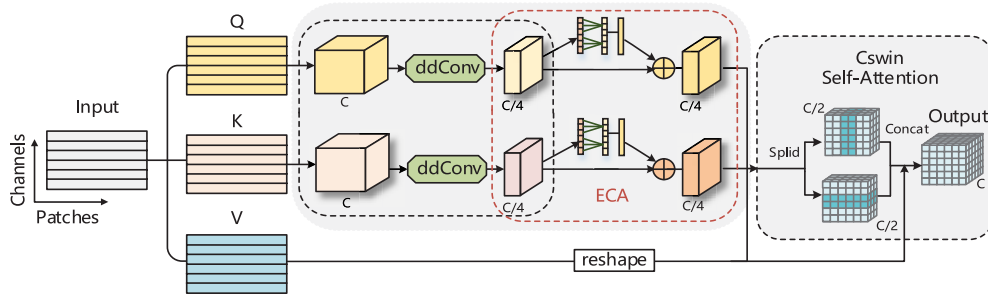


Fig. 4. Window Self-Attention Enhance block.

In the formula, δ represents the sigmoid activation function and GAP denotes global average pooling. Vector x is obtained by reducing the dimensions of $Q \in R^{C/4 \times H \times L}$ and $K \in R^{C/4 \times H \times L}$, which are then subject to global average pooling to obtain dimensions $Q \in R^{C/4 \times 1 \times 1}$ and $K \in R^{C/4 \times 1 \times 1}$. These are compressed and reshaped into dimensions $Q \in R^{C/4 \times 1}$ and $K \in R^{C/4 \times 1}$, followed by a one-dimensional convolution with a kernel size of 3 and activated by the sigmoid function to obtain the channel weight vector Y . Y is then reshaped into a three-dimensional attention vector, which is fused with x to obtain the feature-enhanced Q and K after feature selection. Compared to the original SENet [34], which uses two fully connected operations to activate the attention weight vector, this approach models the local interaction relationship of the channel weight vector using one-dimensional convolution after global average pooling of the input feature map x ($Q \in R^{C/4 \times H \times L}$ and $K \in R^{C/4 \times H \times L}$), effectively reducing the computational and parameter complexity introduced by the model.

After dimensionality reduction and information enhancement, the feature map is input into the CSwin Self-Attention part based on windowed self-attention mechanism for processing. The

CSwin Self-Attention is a representative work of window-based self-attention mechanism, which splits the feature map into two parts along the channel dimension and then extracts the feature information by globally self-attention constrained within horizontal and vertical rectangular windows respectively. The formula of CSwin Self-Attention is as follows:

$$CSwinSelf - Attention(X) = Concat(head_1, \dots, head_K)W$$

$$where\ head = \begin{cases} H - Attention_k(X) & k = 1, \dots, \frac{K}{2} \\ V - Attention_k(X) & k = \frac{K}{2} + 1, \dots, K. \end{cases} \quad (5)$$

where X represents the input feature map. W represents the output calculation matrix, and CSwin Self-Attention represents the process of calculating the self-attention weights on the vertical and horizontal rectangular windows and reassembling the two parallel layers.

3.2.2. Multi-scale group convolution block

The WSE block is responsible for mining global information, but it overlooks the fine-grained feature information within patches. Additionally, capturing multi-scale features is equally important for solving complex scale changes in medical image segmentation. To effectively capture local multi-scale information, we propose a multi-scale group convolution structure named the MGC block. This block is responsible for supplementing the feature map calculated by the WSE block with local multi-scale information, allowing the network to explore global context information and then supplement local feature information. The structure of the MGC block is shown in Fig. 5, where green, red, and yellow lines connecting different-colored blocks represent inputting the channel dimension into different-sized and different-quantity convolution kernels for processing, effectively combining the multi-scale mechanism with group convolution operation.

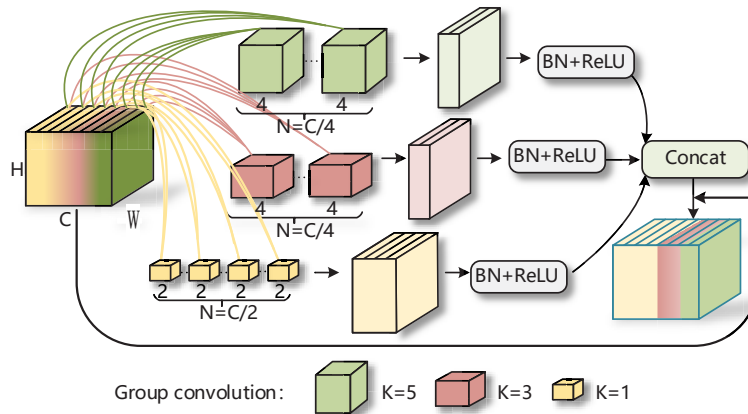


Fig. 5. MGC block structure diagram.

The MGC block consists of three convolution branches and employs the group convolution operation to limit the computation and parameters. The kernel size of the convolutions are 5×5 , 3×3 and 1×1 , respectively, and padding is set to 2, 1, 0 to keep the input and output feature sizes unchanged. Furthermore, to further reduce parameters and computation while maintaining the same output channel size, the output channels of the three branches are set to $C/4$, $C/4$, and $C/2$, respectively, and the outputs of the three branches are concatenated along the channel dimension.

to obtain a feature map with the same channel dimension as the input feature map. Specifically, for the input feature map $X \in \mathbb{R}^{C \times H \times W}$, the calculation process of the MGC is as follows:

$$X_{out} = \text{Concat}(X_1, X_2, X_3) + X. \quad (6)$$

$$X_i = \text{ReLU}(\text{BN}(\text{GroupConv}(X))), i = 1, 2, 3. \quad (7)$$

where X_{out} represents the output feature map. Concat is the concatenation operation performed along the channel dimension. $X_1 \in \mathbb{R}^{\frac{C}{4} \times H \times W}$, $X_2 \in \mathbb{R}^{\frac{C}{4} \times H \times W}$, and $X_3 \in \mathbb{R}^{\frac{C}{4} \times H \times W}$ are the output feature maps of the three group convolution paths, BN represents the batch normalization operation, and ReLU represents the activation function.

The MGC block builds a pyramid structure from the perspective of the receptive field and the number of convolution kernels, from top to bottom. This pyramid structure captures multiscale local information by using three parallel branches to process spatial local features mined by different-sized receptive fields. In our experimental study, we found that using a limited number of group convolutions instead of ordinary convolutions can maintain the computation accuracy of the network while reducing the number of parameters and computational costs by several times. Therefore, we apply different group convolutions to the three parallel branches to compress the computational costs and balance the outputted feature information by increasing the channel number of the outputted feature maps in each branch with the decrease of the receptive field of the convolution kernels. Finally, the information from different branches is aggregated in the channel dimension. Additionally, we add a residual connection after the channel information fusion to mitigate the semantic gap problem that may occur before and after information aggregation.

4. Experiments

4.1. Datasets and experimental details

In order to verify the effectiveness of the proposed model, a series of ablation experiments and comparison experiments were conducted on the ISIC-2018 challenge dataset and the MICCAI 2017 LiTS dataset. We randomly split the two datasets mentioned above into training, validation, and testing sets in an 8:1:1 ratio. All the network models involved in the experiments of this paper were implemented using the Pytorch framework and uniform hyperparameter settings were made. Adam was chosen as the model optimizer. Before the experiments, all images were reshaped to 512, and the maximum number of training epochs was set to 50. The learning rate was set to 0.0001 and halved every three epochs. The batch size was set to 6, and the cross-entropy loss function was used as the loss function. To prevent severe overfitting of the model, the training of the model was stopped when the model did not save weights for ten consecutive rounds. All experiments were conducted on a machine with 24GB memory and a GeForce RTX 3090.

4.2. Evaluation metrics

To evaluate the performance of the network proposed in this study, the following indicators were used as the evaluation criteria:

$$DICE = \frac{2 \times |A| \cap |B|}{|A| + |B|} \quad (8)$$

$$IOU = \frac{|A \cap B|}{|A \cup B|} \quad (9)$$

$$SEN = \frac{|A| \cap |B|}{|B|} \quad (10)$$

where A and B represent the number of pixels with a value of 1 in the predicted label image and that in the true label image, respectively. Sensitivity (SEN) $\in (0, 1)$, $DICE \in (0, 1)$, intersection over union (IOU) $\in (0, 1)$, the closer the SEN , $DICE$ and IOU are to 1, the better the segmentation performance.

4.3. Ablation studies

In order to analyze the proposed TCI module in depth, the experiments are conducted using UNet as the baseline comparison network, gradually adding the WSE block and MGC block. The specific methods are as follows: (1) Replace the original UNet's convolution layer with the unmodified CSwin-Transformer to form a CSwin-UNet network constructed purely with the Transformer structure. (2) Replace CSwin-UNet with the proposed WSE block to obtain the WSE-UNet network. (3) Based on the previous step, add the MGC block to supplement the WSE block with multi-scale local information to obtain TCI-UNet. The experimental results, as shown in Table 1, demonstrate the importance of windowed self-attention mechanism in medical image segmentation tasks with improved performance of CSwin-UNet on LiTS and ISIC datasets. The improved performance of WSE-UNet, after replacing it with the WSE block, further proves the effectiveness of dimension reduction and information enhancement. The TCI-UNet, obtained by adding the multi-scale group convolution module, once again achieved performance improvement based on the former, proving that the combination of multi-scale thinking and windowed self-attention mechanism can further enhance network performance and also reflects the importance of interaction between local information and global information. Furthermore, to validate the effectiveness of the WSE module, we conducted tests on networks progressively augmented with dimensionality reduction convolution (CSwin-UNet + ddc) and ECA (CSwin-UNet + ECA) based on the CSwin-UNet. This is demonstrated in Table 1, rows four and five. It can be observed that the addition of dimensionality reduction convolution to the CSwin-UNet network led to a mere decrease of 0.30% and 0.47% in the MIOU metrics on the two datasets. Despite this, there was a noticeable enhancement in the network's computational speed (as depicted in Fig. 6), demonstrating that the dimensionality reduction convolution can effectively reduce the model's computational complexity, thereby improving its operational efficiency. On the LiTS dataset, although the CSwin-UNet network did not achieve further segmentation performance improvement compared to the UNet network, the introduction of an ECA-enhanced U-shaped network within the CSwin architecture significantly boosted segmentation performance. This indicates that the incorporation of ECA effectively strengthens the interaction among feature channels within the CSwin network, subsequently elevating the model's feature expression capacity and enhancing network performance. Building upon this foundation, the integration of dimensionality reduction convolution and ECA in the constructed U-shaped network showcased stable and outstanding segmentation performance on both datasets, while ensuring model efficiency.

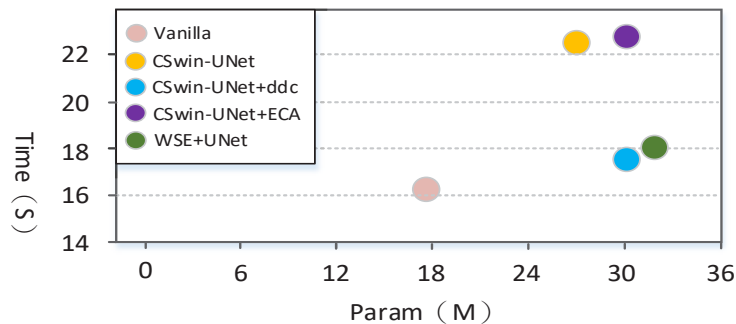


Fig. 6. Comparison chart of network parameter size and computation time.

In order to more intuitively reflect the effect of TCI module on the improvement of segmentation performance, we use the Grad-CAM method to visualize the thermal diagram changes during the construction of TCI module. As shown in Fig. 7, red is the area that the network focuses

Table 1. Ablation Study

Methods	Params (M)	LiTS			ISIC		
		DICE(%)	IOU(%)	SEN(%)	DICE(%)	IOU(%)	SEN(%)
Vanilla	17.267	92.29	88.28	92.03	83.41	74.36	85.23
CSwin-UNet	27.188	92.64	88.09	93.40	85.77	76.69	87.55
CSwin-UNet + ddc	30.387	91.77	87.79	92.25	85.53	76.22	86.81
CSwin-UNet + ECA	30.714	93.14	88.71	93.96	87.41	78.89	88.79
WSE-UNet	31.209	93.28	88.68	93.81	87.37	78.94	88.62
TCI-UNet	33.408	93.81	89.53	95.17	88.22	80.93	90.11

on. The darker the red is, the more attention the network pays to this area. On the contrary, the darker the blue is, the more attention the network ignores this area. In order to further reflect the performance of TCI module, we selected images with greater difficulty in segmentation in two datasets for visual display. The first image in Fig. 7 is an example image of ISIC-2018 dataset. The target area of the image is small, and the contrast between the lesion area and normal skin tissue is very low. The second is a sample image of LiTS-2017 dataset. In this image, the liver region is small, and other tissues account for a large proportion. The gray value of the target region is close to other tissues, which is prone to mis-segmentation. Both images are typical images that are easy to miss diagnosis and misjudge the lesion area in manual film reading. In the case of skin lesion segmentation, UNet showed many false attention regions which led to a high error rate of segmentation. After replacing the convolutional layers in UNet with CSwin-Transformers and constructing a CSwin-UNet, the background noise was significantly reduced, and the network's focus gradually shifted towards the lesion location, while reducing misfocus on irrelevant regions. When the CSwin-UNet was further replaced with WSE blocks to construct a WSE-UNet, the background noise was further reduced, and the network's attention became even more focused on the lesion location. After adding an MGC block to the WSE-UNet to construct a TCI-UNet, the level of attention to the lesion area was further improved, with a deeper and more uniform response in the target region. For liver region segmentation, UNet achieved a noticeable response in the target area, but background noise still had a significant impact on the whole region, spreading the network's computational resources. It can be seen that CSwin-UNet significantly reduces background noise, and WSE-UNet further reduces background noise while making the response of the target area more prominent. Finally, with the addition of the MGC block, the response areas of other tissues decreased significantly, focusing the network's attention on the target area as much as possible, improving the final segmentation performance.

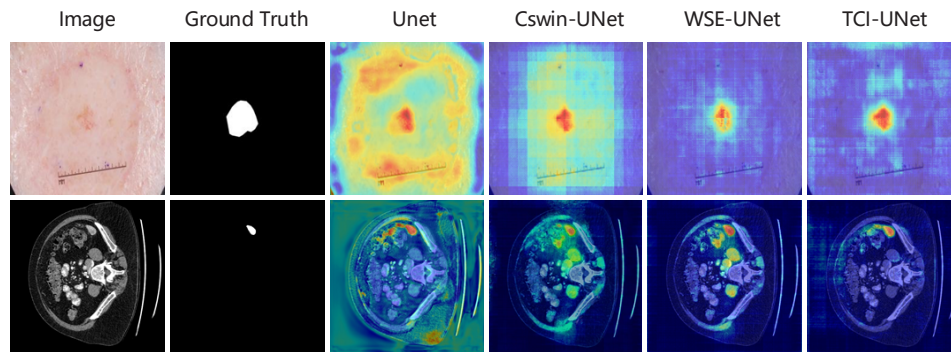


Fig. 7. Comparison of network thermodynamic diagram after gradually adding different blocks.

Table 2. Comparison with state-of-the-art networks

Methods	Years	Para (M)	LiTS			ISIC		
			DICE (%)	IOU (%)	SEN (%)	DICE (%)	IOU (%)	SEN (%)
UNet [2]	2015	17.27	92.18	88.04	91.98	83.50	75.12	85.32
UNet++ [3]	2018	36.63	92.24	88.09	91.77	85.41	76.91	87.16
SwinUNet [14]	2021	27.53	91.60	86.44	91.92	86.45	77.71	87.82
TransUNet [11]	2021	105.9	92.41	88.14	92.44	86.91	78.07	87.68
Transfuse [16]	2021	26.57	93.54	89.39	93.32	86.66	78.03	88.13
ScaleFormer [35]	2022	114.4	91.66	86.98	91.50	86.11	77.58	87.45
UCTransNet [32]	2022	67.23	93.18	88.93	93.22	87.31	78.89	88.43
Transdeeplab [36]	2022	21.16	92.52	87.62	93.27	88.05	79.07	88.55
CE-Net [8]	2019	29.00	92.27	88.18	92.03	87.21	78.82	88.03
CPF-Net [37]	2020	43.27	92.35	88.33	91.77	88.32	80.75	89.07
FAT-Net [12]	2022	28.76	93.46	89.27	94.50	88.14	81.27	90.06
TCI-UNet	33.41	93.92	89.95	94.62	89.52	81.44	91.03	

4.4. Comparisons with other state-of-the-art methods

We compared TCI-UNet with the current state-of-the-art networks, and to comprehensively evaluate the model's performance, we conducted a rigorous five-fold cross-validation. In our experiments, we ensured that all experimental networks were tested under the same computational environment, and detailed experimental results are listed in Table 2. UNet++ and UCTransNet are models improved from the perspective of skip connections. The improvement of DICE index shows that UCTransNet, which introduces the Transformer structure, is more effective in improving the performance of the segmentation network than UNet++, proving the importance of modeling long-range dependencies in the network. UCTransNet and TransUNet combine the Transformer structure with convolution operations outside, and their performance on datasets with many small targets, such as LiTS, is lower than that of TCI-UNet. This proves that TCI module's way of interacting globally and locally inside the network is more friendly to small target segmentation detection and has stronger generalization. Furthermore, compared to TransUNet's 105.9 M parameters and UCTransNet's 67.23 M parameters, TCI-UNet has a significant advantage in the number of parameters, while ensuring segmentation accuracy, it minimizes the waste of computational resources and achieving a good balance between performance and parameters, which is beneficial for model deployment optimization. Furthermore, the results vividly demonstrate the relative superiority of FAT-Net and TCI-UNet in segmentation performance on both datasets. This provides strong support for our conclusions once again, namely, the unique advantage of the Transformer's computational mechanism in capturing long-range dependencies, significantly enhancing the accuracy of lesion area identification from a global perspective. It is worth noting that in the five-fold cross-validation on both datasets, TCI-UNet consistently outperformed FAT-Net in segmentation results, highlighting the outstanding capabilities of the WSE and MGC blocks in TCI-Net in capturing global contextual information and multiscale local details. The organic fusion of these two blocks and their collaborative interaction within the Transformer framework drive our model to exhibit superior performance in segmentation tasks compared to other competitive models. This further validates the outstanding performance of TCI-UNet in terms of accuracy, stability, and generalization.

Figure 8 shows the parameter count and computation time of various networks in the comparative experiments. The horizontal axis represents the parameter count of the network, and the vertical axis represents the computation time. The experiment used two datasets, ISIC-2018

and LiTS-2017, and compared the average computation time of each network on both datasets. From the figure, it can be seen that, compared to most segmentation networks, TCI-UNet has fewer parameters and computation time, achieving a good balance between accuracy and speed. Under the same parameter size, the computation time of TCI-UNet is relatively less than FAT-Net, which reflects the efficient advantage of TCI-UNet in network operation. In practical applications, this efficiency can help us train and infer models faster, thereby improving the performance of the entire system.

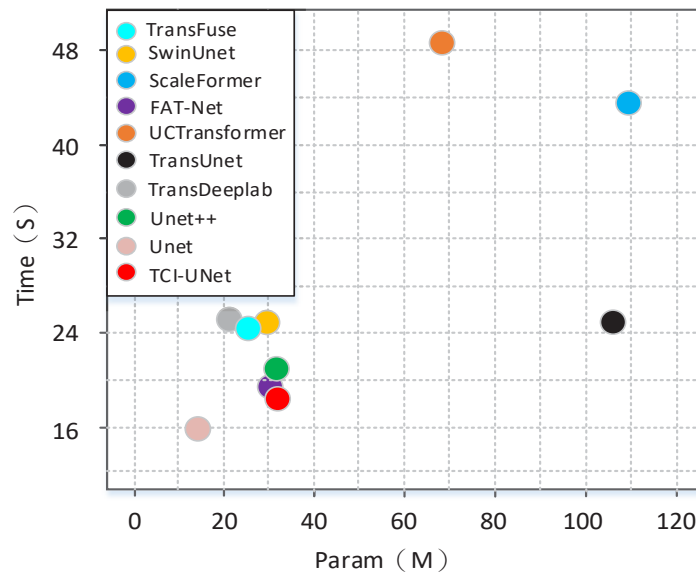


Fig. 8. Comparison chart of network parameter size and computation time.

Figure 9 shows the segmentation comparison results of different networks on the ISIC-2018 dataset. In the comparative experiments of state-of-the-art (SOTA) networks, we selected four representative networks for segmentation comparison. Even when the contrast between the target area and the surrounding tissue is obvious (first row), UNet and UNet++ still have serious over-segmentation issues. The three networks with Transformer mechanism, UCTransNet and TransUNet, and TCI-UNet proposed in this paper greatly improved the over-segmentation problem. Correspondingly, when the contrast between the target area and the surrounding tissue is low (second row), UNet and UNet++ have serious under-segmentation issues, but the three networks with Transformer mechanism still achieved excellent segmentation results, proving the importance of extracting global context information for medical image segmentation. Compared to large parameter networks like UCTransNet and TransUNet, TCI-UNet achieved better segmentation results on skin lesion images, proving the effectiveness of supplementing multi-scale local information for improving segmentation performance.

Figure 10 shows the segmentation comparison results of different networks on the LiTS-2017 dataset. In the comparative experiments of state-of-the-art (SOTA) networks, we selected four networks with the best performance indicators for segmentation comparison. Compared to skin lesion images, liver CT images have small color changes and similar gray-scale values between different tissues. To more clearly show the ability of the TCI module to extract features, two images with clear boundaries between liver tissue and other tissues were selected. By observing the segmentation results in Fig. 10, it can be seen that the TCI-UNet has a smoother edge segmentation for the target area and more precise segmentation for the detailed tissue parts within

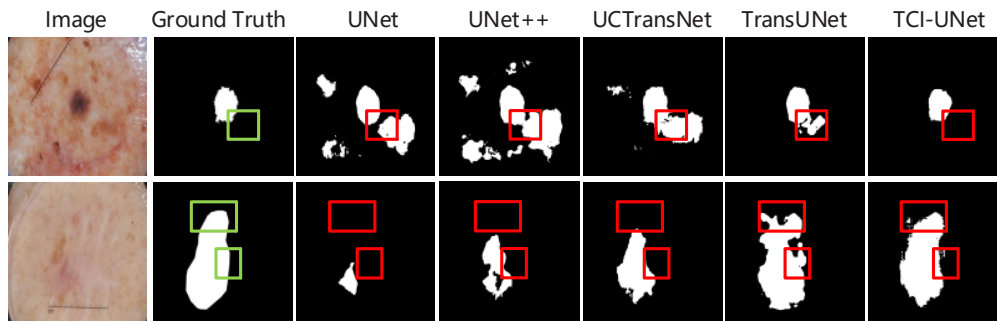


Fig. 9. Comparison of segmentation results on ISIC-2018 dataset.

the target area, further proving the correctness and effectiveness of the idea of combining global context information and local information.

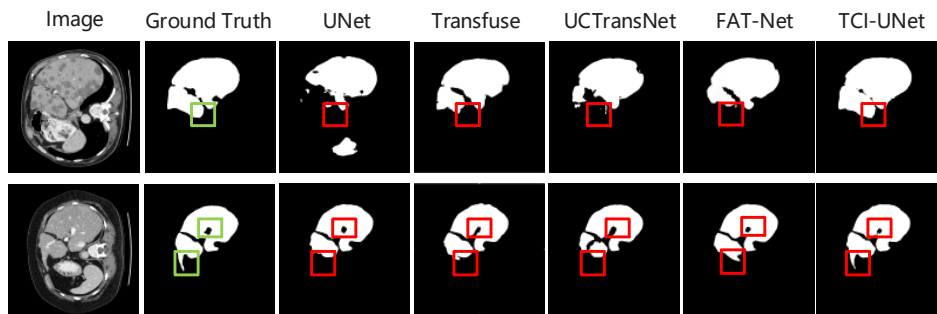


Fig. 10. Comparison of segmentation results on MICCAI 2017 LiTS dataset.

Figure 11 shows box plots of DICE and IOU metrics on the LiTS and ISIC datasets, reflecting the stability and robustness of the five networks tested. Our TCI-UNet achieved the best results in terms of highest, lowest and median values on both datasets. On the LiTS dataset, TransUNet suffered from severe overfitting due to its large number of parameters, showing the importance of

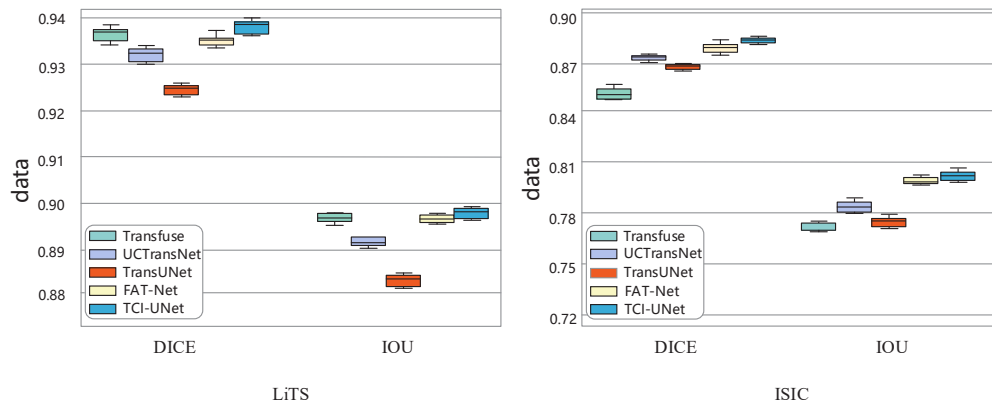


Fig. 11. Boxplot of DICE and IOU indicators on LiTS and ISIC datasets.

controlling model parameters and the superiority of TCI-UNet in this aspect. On the ISIC dataset, our TCI-UNet demonstrated a significant accuracy advantage while maintaining a relatively lightweight design. This fully demonstrates the outstanding competitiveness of TCI-UNet in medical image segmentation tasks.

5. Discussion

We demonstrate the correctness of combining Convolutional Neural Networks and Transformer through experiments, develop a way of interweaving global and local information within the Transformer, and unify the advantages of window-based self-attention mechanism, multiscale thinking, and the combination of global and local information. The experimental results show the superiority of the TCI feature extraction module proposed in this paper, providing an important reference for the combination of Transformer and Convolutional Neural Networks in the field of medical image segmentation. In designing the TCI feature extraction module, we also explored a parallel structure, but the experimental results were not ideal, and it is speculated that the parallel addition of multiscale local information may damage feature integrity when filling in global information, so a serial structure was eventually adjusted. Additionally, when designing the MGC block and conducting experiments, the three branches outputting C-channel feature maps and those outputting C/3-channel feature maps showed similar performance improvements, but the latter significantly reduced the introduced parameters and computation. Based on this, reducing the number of output channels of large convolution kernels balances the feature information quantity of different paths. Finally, through a large number of experiments, it was determined that the best balance between parameter quantity and segmentation accuracy was achieved when the output channel ratios of the three branch convolution kernels were C/2, C/4, C/4, providing meaningful reference for the design of subsequent multi-path information processing work.

6. Conclusion

Medical image segmentation is a critical step in clinical diagnosis and analysis. In this work, we implemented the interaction of global information and multi-scale local information within the Transformer. We also designed a TCI-Net using an U-shape architecture to provide precise and reliable medical image automatic segmentation. By combining end-to-end WSE block and MGC block, the proposed method significantly improves the latest level of medical image segmentation on the ISIC-2018 and LiTS-2017 datasets, demonstrating the advantage of the TCI-Net model. However, there are some limitations in the model as well. Although TCI-UNet has attempted to minimize the number of parameters, the network built on the Transformer structure has relatively high parameters overall. There is still room for improvement in parameter reduction. Moreover, experiments show that the accuracy of TCI-UNet is not stable and there is room for further improvement. In future research, we will address these issues to make TCI-UNet applicable to more medical image segmentation tasks.

Funding. Hebei Provincial Natural Science Fund Key Project (F2017201222); National Natural Science Foundation of China (61473112).

Disclosures. The authors declare there is no conflict of interest.

Data availability. The data presented in this study are available on request from the corresponding author.

References

1. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3431–3440.
2. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, (Springer, 2015), pp. 234–241.

3. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, T. Nima, and L. Jianming, "Unet++: A nested u-net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer 2018), pp. 3–11.
4. H. Huang, L. Lin, R. Tong, Hongjie Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. Chen, and J. Wu, "UNet 3+: A full-scale connected unet for medical image segmentation," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, 2020), pp. 1055–1059.
5. X. Li, H. Chen, X. Qi, Q. Dou, C. Fu, and P. Heng, "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imaging* **37**(12), 2663–2674 (2018).
6. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4700–4708.
7. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *arXiv*, arXiv:1804.03999 (2018).
8. Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "CE-Net: Context encoder network for 2d medical image segmentation," *IEEE Trans. Med. Imaging* **38**(10), 2281–2292 (2019).
9. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*(2017), pp. 5998–6008.
10. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," *arXiv*, arXiv:2010.11929 (2020).
11. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transformers make strong encoders for medical image segmentation," *arXiv*, arXiv:2102.04306 (2021).
12. H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.* **76**, 102327 (2022).
13. Y. Zhang, H. Liu, and Q. Hu, "Fusing transformers and cnns for medical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, Cham, 2021), pp. 14–24.
14. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10012–10022.
15. A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 12894–12904.
16. X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp.12124–12134.
17. A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 6185–6194.
18. Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "VSA: learning varied-size window attention in vision transformers," *arXiv*, arXiv:2204.08446 (2022).
19. T. Yu, G. Zhao, P. Li, and Y. Yu, "BOAT: bilateral local attention vision transformer," *arXiv*, arXiv:2201.13027 (2022).
20. C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang, and S. Yan, "Inception transformer," *arXiv*, arXiv:2205.12956 (2022).
21. Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7287–7296, (2022).
22. H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.6824–6835 (2021).
23. X. Qin, Z. Zhang, and C. Huang, *et al.*, "Going deeper with nested U-structure for salient object detection," *Pattern recognition* **106**, 107404 (2020).
24. M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *arXiv*, arXiv:1802.06955 (2018).
25. N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural networks* **121**, 74–87 (2020).
26. W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Med. Phys.* **46**(2), 576–589 (2019).
27. Z. Ni, G. Bian, X. Zhou, Z. Hou, X. Xie, C. Wang, Y. Zhou, R. Li, and Z. Li, "Residual attention u-net for semantic segmentation of cataract surgical instruments," *arXiv*, arXiv:1909.10360 (2019).
28. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Jiang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," *arXiv*, arXiv:2105.05537 (2021).
29. A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Lu, "DS-TransUNet: Dual swin transformer u-net for medical image segmentation," *arXiv*, arXiv:2106.06716 (2021).
30. B. Chen, Y. Liu, Z. Zhang, G. Lu, and AWK. Kong, "Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation," *arXiv*, arXiv:2107.05274 (2021).

31. Y. Sha, Y. Zhang, X. Ji, and L. Hu, "Transformer-Unet: Raw Image Processing with Unet," [arXiv](#), arXiv:2109.08417 (2021).
32. H. Wang, P. Cao, J. Wang, and O.R. Zaiane, "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*.36(3), pp.2441–2449 (2022).
33. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net:Efficient channel attention for deep convolutional neural networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 11534–11542.
34. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp.7132–7141.
35. H. Huang, S. Xie, L. Lin, Y. Iwamoto, X. Han, Y. Chen, and R. Tong, "ScaleFormer: revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation," [arXiv](#), arXiv:2207.14552 (2022).
36. R. Azad, M. Heidari, M. Shariatnia, E.K. Aghdam, S. Karimijafarbigloo, E. Adeli, and D. Merhof, "TransDeepLab: Convolution-free transformer-based deeplab v3 + for medical image segmentation," [arXiv](#), arXiv:2208.00713 (2022).
37. S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "CPFNet: context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imaging* **39**(10), 3008–3018 (2020).